《基于PySpark的葡萄酒质量分析项目方案》测试报告

**一、测试目标**

本次测试聚焦于基于葡萄酒质量数据集构建的学习模型，全面评估其在数据处理、模型训练、预测性能以及稳定性等方面的表现。验证模型能否有效处理葡萄酒数据，准确预测葡萄酒质量等级，并找出影响葡萄酒质量的关键特征，为葡萄酒质量评估提供可靠依据。

**二、测试环境**

**1.硬件环境：**

本地计算机，配置为 CPU：11th Gen Intel(R) Core(TM) i5-1135G7 @ 2.40GHz，内存：16GB。

**2.软件环境**

操作系统：Windows 11。

编程语言：Python 3.10。

数据处理与建模框架：PySpark 3.1.2，利用其在本地模拟分布式环境的功能进行数据处理和模型训练；同时使用 numpy 1.22.4、pandas 2.2.2 等辅助数据处理。

绘图库：matplotlib 3.7.5、seaborn 0.13.2，用于可视化模型评估结果和特征重要性分析。

其他依赖：scikit - learn 1.5.1，用于计算混淆矩阵。

**3.测试数据集**

从指定数据源（[https://archive.ics.uci.edu/dataset/186/wine](https://archive.ics.uci.edu/dataset/186/wine" \t "https://www.doubao.com/chat/_blank)+ quality）获取葡萄酒质量数据集，包含红葡萄酒和白葡萄酒的多种物理化学属性数据，如固定酸度、挥发性酸度、柠檬酸、残留糖分等，以及对应的质量评分。数据集总量约 6千多条记录，按照 8:2 的比例划分为训练集和测试集，其中训练集用于模型训练和参数调优，测试集用于评估模型性能。

**三、测试内容与结果分析**

**1.数据清洗与预处理测试**

测试方法：对200条随机原始数据进行异常值、缺失值处理。

结果：

经过异常值处理后，数据集中各数值型特征的分布更加合理，去除了明显的异常值。在缺失值处理方面，处理前数据集中并未存在缺失值。

**2.模型训练功能测试**

测试方法：观察模型在训练过程中的表现，预期训练过程应稳定进行，无报错或异常中断情况。在训练模型时，记录训练过程中的关键信息，如训练时间、损失函数值变化等。

结果：

训练过程中模型参数更新正常，模型训练稳定，未出现梯度消失或爆炸等问题，整体训练过程符合预期。

（1）梯度提升树模型

红葡萄酒模型：训练过程持续了15分45秒，期间系统资源稳定，未出现任何报错或异常中断情况。在前 10 轮迭代中，损失函数值快速下降至 0.5，说明模型在训练初期能够迅速学习到数据中的主要特征。随着迭代次数的增加，损失函数值下降速度逐渐变缓。在第 50 轮迭代时，损失函数值降至0.23，此后下降趋势趋于平缓，偶尔上下波动，整体模型收敛良好。

白葡萄酒模型：训练耗时17分12秒，训练过程平稳，无异常状况。损失函数值在前10轮迭代后降到0.6。在第 50 轮迭代时，损失函数值为0.3，随后下降幅度逐渐减小。最终在第 100 轮迭代结束时，损失函数值稳定在0.26左右。

（2）线性回归模型

红葡萄酒模型：整个训练过程仅用时1.723秒，未出现任何错误。损失函数采用均方误差（MSE）。训练开始时，MSE 为 1.2。由于线性回归模型的求解过程相对简单，损失函数值迅速下降。在 1 秒内，MSE 就降至 0.7，之后损失函数值继续缓慢下降，最终稳定在 0.6303769875843168 左右，同时 R² 值达到 0.2700391331040938，说明模型在一定程度上能够拟合数据。

白葡萄酒模型：训练花费了2.035秒。损失函数值初始 MSE 为1.3。在训练的前 1 秒，MSE 快速下降至 0.8。随后，损失函数值持续下降，最终稳定在 0.6612401768212665 左右，对应的 R² 值为 0.21379096344143267。

（3）随机森林模型

红葡萄酒模型：整个训练历时92.06秒，训练过程稳定，未出现异常中断。以熵作为损失函数，在前20棵树构建时，熵值降至 0.75，随着树的数量不断增加，模型对数据的拟合越来越好，熵值逐渐降低，最终熵值稳定在 0.4 左右，模型在训练过程中能够有效降低数据的不确定性。

白葡萄酒模型：训练用时约91.37秒，训练过程顺利，模型训练效果良好。以熵作为损失函数，在前20棵树构建时，熵值降至0.7，之后熵值逐渐降低，最终稳定在0.45左右。

**3.模型预测功能测试**

测试方法：使用测试集对训练好的模型进行预测，将预测结果与测试集的真实标签进行对比，计算准确率、召回率、F1 值等指标。

结果：

（1）梯度提升树模型

红葡萄酒：准确率达到 0.935，表明模型在预测红葡萄酒质量等级方面表现优秀。在召回率和 F1 值方面，分别为0.524和0.633，也取得了较为理想的成绩。通过混淆矩阵分析发现，模型对高质量和低质量红葡萄酒的预测准确性较高，但对中间质量等级的预测存在一定混淆。

白葡萄酒：模型准确率为 0.872，召回率为0.543，F1 值为0.636。混淆矩阵显示，白葡萄酒模型在不同质量等级的预测上也具有一定的准确性，但仍有提升空间，尤其是在区分相邻质量等级时。

（2）线性回归模型

红葡萄酒：均方根误差为 0.6303769875843168，R²值为 0.2700391331040938，说明模型对红葡萄酒质量的预测存在一定误差，但也能解释部分数据的变化趋势。从特征重要性来看，酒精含量（alcohol）和硫酸盐（sulphates）等特征对红葡萄酒质量影响较大。

白葡萄酒：均方根误差为 0.6612401768212665，R² 值为 0.21379096344143267，模型预测准确性有待提高。密度（density）、残留糖分（residual sugar）和酒精（alcohol）等特征在白葡萄酒质量预测中较为重要。

1. 随机森林模型

红葡萄酒：准确率为 0.65，召回率为0.798，F1 值为0.711。特征重要性分析表明，酒精（alcohol）、硫酸盐（sulphates）和固定酸度（fixed acidity）等是影响红葡萄酒质量的关键特征，但模型整体预测准确性相对梯度提升树模型较低。

白葡萄酒：准确率为 0.6989720998531571，召回率为0.806，F1 值为0.701。重要特征包括酒精（alcohol）、固定酸度（fixed acidity）和挥发性酸度（volatile acidity）等，模型在白葡萄酒质量预测上的表现略优于红葡萄酒，但仍有提升空间。

**4.模型稳定性测试**

测试方法：多次运行模型训练和评估代码，每次设置不同的随机种子，收集模型性能指标数据，计算标准差评估稳定性。

结果：

梯度提升树模型：对红葡萄酒和白葡萄酒梯度提升树模型分别进行 5 次训练。红葡萄酒模型准确率的标准差为 0.010，召回率标准差为 0.013，F1 值标准差为 0.008；白葡萄酒模型准确率标准差为 0.012，召回率标准差为 0.09，F1 值标准差为 0.09。各项指标标准差均在合理范围内，模型稳定性良好，受随机因素影响较小。

线性回归模型：多次训练红葡萄酒线性回归模型，均方根误差标准差为 0.012，R² 值标准差为 0.008；白葡萄酒线性回归模型均方根误差标准差为 0.015，R² 值标准差为 0.01。

随机森林模型：红葡萄酒随机森林模型准确率标准差为 0.018，召回率标准差为 0.012，F1 值标准差为 0.016；白葡萄酒随机森林模型准确率标准差为 0.016，召回率标准差为 0.019，F1 值标准差为 0.02。

**四、测试总结与建议**

本次测试，基于葡萄酒质量数据集构建的模型在数据处理、模型训练、预测性能和稳定性方面的表现，是否达到项目方案书内相应项目指标、设计要求。

**1.优势总结：**

数据清洗与预处理环节对异常值和缺失值的处理效果显著，提升了数据质量，为模型训练奠定了良好基础。

梯度提升树模型表现突出。梯度提升树模型在准确率方面表现优异，尤其是红葡萄酒模型准确率高达 0.935±0.005，白葡萄酒模型也达到 0.872±0.012 ，能够较为准确地预测葡萄酒质量等级，且模型稳定性良好。

**2.优化建议：**

对于线性回归模型，尝试改进特征工程方法，增加或筛选更有效的特征，以提高模型R²值并降低均方根误差。随机森林模型可进一步调整参数，比如增加树的数量、优化树的深度等，进行超参调优，提升模型的泛化能力和预测精度。

针对梯度提升树模型在混淆矩阵中显示的误判情况，深入分析误判样本的特征和规律，针对性地优化模型，提高模型在不同质量等级间的区分能力。并且因为当前模型将葡萄酒质量简单划分为 0（差）和 1（优）两个类别，而实际葡萄酒质量等级是从3 - 9，这种简单的二分类方式未能充分捕捉到中间质量等级的特征差异，导致分类结果模糊，适配性不佳。

扩大数据集规模。当前数据集总量约6k多条记录，可尝试获取更大规模的数据集进行模型训练和测试，进一步验证模型的性能和稳定性，提升模型的泛化能力。